АДАПТИВНЫЙ ИНТЕРФЕЙС УТОЧНЕНИЯ ЗАПРОСОВ К СИСТЕМЕ КОНТЕНТ-МОНИТОРИНГА INFOSTREAM

ADAPTIVE REQUEST PRECISION INTERFACE IN CONTENT-MONITORING SYSTEM INFOSTREAM

А.Н. Григорьев, gri@visti.net, Д.В. Ландэ, dwl@visti.net

Описывается адаптивный механизма уточнения запросов в информационно-поисковой системе InfoStream. На основе методов кластерного анализа создан пользовательский интерфейс, учитывающий различные аспекты представления документов, входящих в базы данных системы

Интерфейсы режимов «расширенного поиска» ведущих поисковых систем в Internet порой вызывают уныние даже у продвинутых пользователей. С одной стороны, открывается многообразие возможностей, а с другой – логическое объединение значений заполненных полей зачастую приводит к нулевым результатам (при том, что основная проблема Internet как информационной среды – это избыточность информации).

В связи с этим в последнее время получили распространение адаптивные интерфейсы уточнения запросов, чаще всего реализуемые путем кластеризации результатов первичного поиска. Появилось такое понятие, как метод "папок поиска" (Custom Search Folders), который не связывается с определенным алгоритмом кластеризации, а представляет собой множество подходов, общее у которых - попытка сгруппировать результаты поиска и представить кластеры в удобном для пользователей виде.

К подобным механизмам можно отнести, например, австралийский поисковый сервер Mooter (http://www.mooter.com), на котором применяется визуальный подход к предоставлению результатов поиска по обрабатываемым запросам путем группировки результатов первичного поиска по категориям. Другой поисковый сервер iBoogie (http://www.iboogie.com/) также группирует результаты поиска, но отображает их в виде, близком к экрану проводника Windows. Слова и словосочетания в информационных портретах, применяемых, например, в системе Галактика Зум, также позволяют адаптивно уточнять первичные запросы.

В Информационном центре «ЭЛВИСТИ» была разработана система контент-мониторинга InfoStream, которая применяется для решения задач автоматизированного сбора новостной информации с открытых web-сайтов, ее обработки, систематизации и обеспечения доступа к ней в поисковых режимах. Эта система в настоящее время охватывает свыше 1200 источников – более 30000 уникальных новостных сообщений в сутки, при этом в архивах (ретроспективных базах данных) системы хранится свыше 20 млн. записей.

Поисковым механизмом InfoStream является полнотекстовая ИПС InfoReS, обладающая мощным языком запросов. При определенном уровне наполнения баз данных системы понадобился механизм уточнения запросов, доступный не только профессионалам, но и простым пользователям, в запросах которых среднее количество слов по статистике не превышает двух-трех.

Система контент-мониторинга InfoStream используется в промышленном режиме уже 5 лет. На начальном этапе внедрения системы был реализован принцип автоматической рубрикации сообщений, использования предопределенных администратором запросов. В дальнейшем были подключены возможности уточнения запросов наиболее весомыми по эмпирико-статистическим критериям ключевыми словами, входящими в результаты первичного поиска. Однако данные подходы еще не позволяли уточнять запросы с использованием многочисленных параметров, присутствующих в первичных документальных выборках.

Простейшего информационного портрета оказалось мало – понадобился «информационный альбом» - многоаспектная подборка параметров выборки по первоначально составленному запросу. И такая возможность была реализована. При этом в отличие от большинства подобных систем, в InfoStream уточняющие параметры поиска задаются не заполнением сложной формы расширенного поиска, а указываются путем выбора из информационного альбома, получаемого в результате поиска по первичному запросу.

Конечно, новые возможности потребовали существенного пересмотра концепции индексирования документов, выбора из текстов документов и нормализации ключевых

слов-дескрипторов, выявления ряда содержательных параметров документов.

Сегодня в системе InfoStream информационный альбом (Рис.1), соответствующий первичному запросу, содержит такие параметры, как ключевые слова, рубрики, языки, страны, размеры документов.

В частности, в адаптивном интерфейсе системы существенно облегчен множественный выбор источников информации, соответствующих заданному запросу.

Пользователю системы InfoStream доступна возможность задания характеристик размеров искомых документов. Это может быть использовано, например, как при поиске объемных аналитических материалов, обзоров, законодательных актов, так и при поиске кратких резюме или сводок. В системе предусмотрен выбор трех уровней размеров сообщений: высокий (leng.large) - свыше 10000 символов, средний (leng.medium) - свыше 1000 символов и низкий (leng.small) - до 1000 символов.

Предусмотрен и такой «экзотический» параметр, как уровень насыщенности документов цифровой информацией. Эта возможность полезна, например, при поиске аналитических документов, ценовых таблиц, рейтингов и т.п. В системе выделено три уровня насыщенности документов цифровой информацией: высокая (numb.large) - свыше 10%, средняя (numb.medium) - свыше 3% и низкая (numb.small) - до 3%.

Именно адаптивность делает интерфейс удобным для пользователя, который по запросу «Нацбанк Украины» не будет анализировать массив из 1200 web-сайтов, а выберет, например 5-6 наиболее актуальных из 42 реально отражающих это понятие источников. Одновременно запросив высокий уровень насыщенности цифровой информацией, легко выйти на сводку о валютном рынке Украины (Рис.2).

С развитием информационных ресурсов Internet вечная проблема поиска информации сегодня получила новое звучание: "поиск информации в неограниченной неоднородной динамической среде". Поэтому одним из самых перспективных направлений обработки информации в настоящее время является контент-мониторинг - непрерывный процесс анализа текстовых массивов.

В системе InfoStream в настоящее время реализованы алгоритмы автореферирования, построения сюжетов – семантических цепочек документов, таблиц взаимосвязей понятий, динамики их упоминаний. Непрерывная, конвейерная обработка информационных потоков является самой характерной чертой системы, которая нашла широкое применение для поддержки принятия решений в таких областях, как государственное управление, анализ товарных рынков, реклама, маркетинг, поиск партнеров и клиентов, отслеживание деятельности конкурентов.

Благодаря возможностям адаптивного интерфейса уточнения запросов, кластеризации результатов первичного поиска, система InfoStream решает не только задачу поиска необходимой информации, но и вплотную подходит к проблемам обобщения данных и их содержательного анализа.

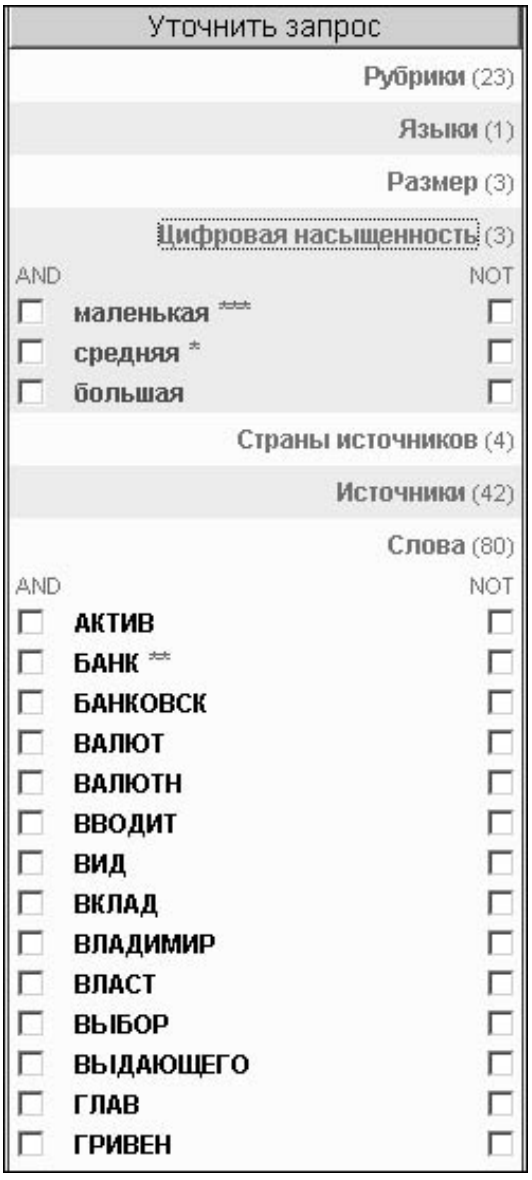

Рис.1. Информационный альбом

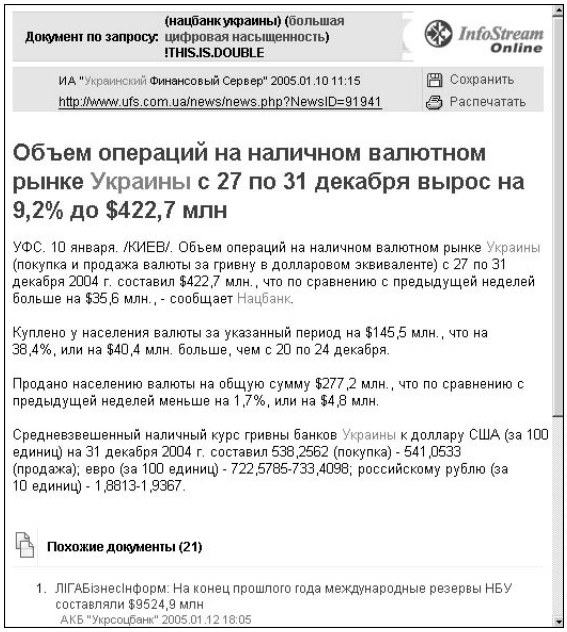

Рис.2. Релевантный документ

Список литературы

1. Кириченко К.М, Герасимов М.Б. Обзор методов кластеризации текстовых документов // Материалы международной конференции Диалог'2001, (http://www.dialog-21.ru/Archive/ 2001/volume2/2_26.htm)
2. Ландэ Д.В. Поиск знаний в Internet. Профессиональная работа // М.: Издательский дом «Вильямс», 2005. –272 с.
3. Ландэ Д.В. Поисковые системы: поле боя – семантика // Киев. Журнал "Телеком", № 4, 2004, С. 44-50